Journal of Electronic Research and Application, 2025, Volume 9, Issue 1
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Building a Diabetes Prediction System Based on Machine Learning Algorithms

Shubo Liang*

Southeast University, Nanjing 210018, Jiangsu Province, China

Corresponding author: Shubo Liang, cuteshubo1997@163.com

Abstract: This paper explores the possibility of using machine learning algorithms to predict type 2 diabetes. We selected two commonly used classification models: random forest and logistic regression, modeled patients' clinical and lifestyle data, and compared their prediction performance. We found that the random forest model achieved the highest accuracy, demonstrated excellent classification results on the test set, and better distinguished between diabetic and non-diabetic patients by the confusion matrix and other evaluation metrics. The support vector machine and logistic regression perform slightly less well but achieve a high level of accuracy. The experimental results validate the effectiveness of the three machine learning algorithms, especially random forest, in the diabetes prediction task and provide useful practical experience for the intelligent prevention and control of chronic diseases. This study promotes the innovation of the diabetes prediction and management model, which is expected to alleviate the pressure on medical resources, reduce the burden of social health care, and improve the prognosis and quality of life of patients. In the future, we can consider expanding the data scale, exploring other machine learning algorithms, and integrating multimodal data to further realize the potential of artificial intelligence (AI) in the field of diabetes.

Keywords: Type 2 diabetes; Machine learning; Predictive modeling; Artificial intelligence; Chronic disease management

Online publication: February 12, 2025

1. Introduction

Diabetes mellitus has been classified by the World Health Organization (WHO) as one of the top four non-communicable diseases (NCDs) requiring the closest attention and has become a major threat to global health [1]. As of the latest statistics, the number of diabetic patients worldwide reached 591 million [2]. With the development of national economies and aging demographics, the number of elderly diabetics is on a continuous rise, which has become a major concern in the medical field [3]. Among these, type 2 diabetes accounts for the vast majority of all diabetes cases, approximately 90%. The development of this disease is closely related to several factors, especially obesity, increasing age, and daily living habits [4]. Therefore, effective prediction of diabetes mellitus is

of great significance, which not only helps patients better manage their blood glucose but also helps to reduce the occurrence of complications, thus improving the quality of life of patients [5].

With the development of mobile Internet, many mobile applications (apps) have emerged in the market to help patients manage diabetes. Although many diabetes management mobile apps currently available in the market provide functions such as dietary records, blood glucose tracking, and exercise programs, most of them have not yet fully utilized the potential of artificial intelligence, especially in predictive analytics. Artificial intelligence, as one of the most cutting-edge technologies today, allows us to optimize the prediction and management of diabetes by using machine learning algorithms.

Machine learning, as an important branch of artificial intelligence, has been widely used in the field of healthcare in recent years. More and more scholars have begun to study the potential of machine learning in health management [6–8]. By learning and training large-scale data, machine learning can automatically extract the complex patterns hidden in the data, thus realizing the prediction of unknown samples. Currently, machine learning has achieved promising results in risk prediction of various diseases such as heart disease and stroke. By introducing machine learning into diabetes prediction, it is expected that the key factors of diabetes can be mined from patients' clinical indicators, lifestyles, and other multi-dimensional data, and a personalized risk assessment model can be constructed. This will not only help doctors screen high-risk groups more accurately but also provide patients with customized health management plans, thus achieving early prevention and timely intervention of diabetes.

However, although machine learning has shown great promise in the medical field, research in the task of diabetes prediction has been relatively limited. Most of the existing work focuses only on the performance of a single machine learning algorithm and lacks a systematic comparison and evaluation of different algorithms. In addition, specific prediction models for type 2 diabetes have to be further developed and optimized. Based on the above background, in this paper, we propose to select three common machine learning algorithms in clinical practice: random forest and logistic regression and conduct an empirical study on a real diabetes dataset to compare their predictive performances through multiple evaluation metrics, such as confusion matrix, and to discuss the advantages and disadvantages of each of them. We expect to deepen the understanding of the application of machine learning algorithms in the prediction of type 2 diabetes mellitus through this study, to provide a reference for subsequent studies, and to contribute to the intelligent management of diabetes mellitus.

The significance of this study is that it is expected to fully utilize artificial intelligence technology to address the growing public health challenge of diabetes. Random forest algorithm's powerful data processing ability and anti-interference capability make it very suitable for the application of complex and variable diabetes-related data. Through this study, we can not only promote the innovation of diabetes prediction and management models but also provide useful practical experience for the intelligent prevention and control of chronic diseases. In the long run, this will help to alleviate the pressure on medical resources, reduce the burden of social health care, and maximize the prognosis and quality of life of diabetic patients.

2. Training diabetes prediction models

2.1. Data preparation

The diabetes prediction dataset used in this study is from the Hugging Face platform, which provides rich machine-learning datasets and pre-trained models [8]. The Hugging Face is a well-known open-source community dedicated

to advancing the fields of natural language processing, computer vision, and the like. It provides convenient tools and resources for researchers and developers.

The dataset contains the following fields: sample number (ID), body mass index (BMI), physical health (PhysHlth), age (Age), whether they have high blood pressure (HighBP), whether they have high cholesterol (HighChol), cholesterol check (CholCheck), whether they smoke (Smoker), whether they have had a stroke (Stroke), whether they have experienced heart disease or a heart attack (HeartDiseaseorAttack), whether they engage in physical activity (PhysActivity), fruit consumption frequency (Fruits), vegetable consumption frequency (Veggies), heavy alcohol consumption (HvyAlcoholConsump), whether they have any healthcare (AnyHealthcare), whether they have not seen a doctor due to cost (NoDocbcCost), general health (GenHlth), mental health (MentHlth), whether they have difficulty walking (DiffWalk), gender (Sex), level of education (Education), income level (Income), and whether they are diabetic, dichotomous label (Diabetes_binary).

To prepare the data, we first preprocess the training and test sets, which include handling missing values, data cleaning, and feature selection. After completing the data preprocessing, we divided the training set and test set into feature variables and target variables, respectively. The feature variable included all columns except the diabetes label, and the target variable was the diabetes label column. To reduce the dimensionality of the data and extract the key information, we used the principal component analysis (PCA) method to reduce the dimensionality of the feature variables. With PCA, we converted the high-dimensional feature space into a low-dimensional feature space, while retaining the main variance information of the data. In this study, we choose the dimension of 8 after dimensionality reduction to balance the information retention and computational efficiency.

After PCA dimensionality reduction, we obtain the processed training features and test features, which will be used for subsequent model training and evaluation. At the same time, in order to facilitate the saving and reuse of the models, we save the trained PCA model and the random forest model into files respectively.

Through the above data preparation steps, we have preprocessed, selected, and downscaled the original data, and obtained high-quality training data and test data, which lays the foundation for the subsequent model training and evaluation [9,10].

2.2. Training methods

After data preparation, we use the random forest algorithm to train and predict the processed data. First, we initialize a random forest classifier and set the relevant hyperparameters. Among them, parameters such as the number of decision trees and random states are appropriately configured to optimize the performance of the model and ensure the reproducibility of the results.

Next, we input the preprocessed training features and training labels into the random forest classifier to train the model. During the training process, the random forest algorithm randomly selects a subset of features and samples to build multiple decision trees. Each decision tree classifies the data independently, and the final output of the integrated model is obtained by voting or averaging. Through this integrated learning approach, the random forest algorithm can effectively reduce the variance of a single decision tree and improve the generalization ability of the model. Simultaneously, since random forest introduces randomness in feature selection and sample selection, it is robust to noise and outliers in the data. After the model training is completed, we save the trained random forest model to a file for subsequent model evaluation and application. By serializing and persisting the model, we can easily load and use the trained model in different environments.

We also used a logistic regression algorithm to train and predict the processed data for comparison. First, we

read the training and test set data and processed specific columns in the data. Then, a logistic regression classifier was initialized and a selected subset of features and corresponding labels were input into the classifier for training. By setting parameters such as random states and a maximum number of iterations, we ensured the repeatability and convergence of the model. After the model training was completed, we used the trained logistic regression model to predict the test set and calculated the accuracy of the model. By comparing the prediction results with the actual labels, we obtained an accuracy of 0.85 for the model on the test set, indicating that the model has good classification performance. To further analyze the performance of the model, we plotted the heat map of the confusion matrix. The confusion matrix shows the predictions of the model on different categories, including the number of true positives, true negatives, false positives, and false negatives. By looking at the confusion matrix, we can evaluate the performance of the model in each category and identify possible false predictions of the model.

3. Results

We evaluated our trained random forest model for diabetes prediction using a test set. First, we preprocessed the test data and reduced its dimensions with PCA. Then, we used the model to make predictions, achieving an accuracy of 86%. This shows that the model can effectively distinguish between diabetic and non-diabetic patients.

To further assess the model, we created a confusion matrix heat map, which displays true positives, true negatives, false positives, and false negatives. The results indicate that most predictions were correct, although there were some misclassifications, such as diabetic patients being incorrectly labeled as non-diabetic and vice versa.

Overall, the random forest model with PCA dimensionality reduction demonstrated high accuracy and good performance in predicting diabetes. The confusion matrix analysis helps identify areas for improvement, confirming the effectiveness of our approach for diabetes prediction and its potential use in medical diagnostics and preventive measures. To further evaluate the performance of different machine learning algorithms on the diabetes prediction task, we also conducted experiments using a logistic regression model. Similar to the random forest model, we divided the preprocessed test data into feature variables and target variables and used the trained PCA model to downscale the test features.

Next, we loaded the pre-trained logistic regression model and input the dimensionality-reduced test features into the model for prediction. By comparing the prediction results with real test labels, we calculated the accuracy of the logistic regression model on the test set. The experimental results show that the logistic regression model achieves an accuracy of 83.33% on the test set, which is slightly lower than the performance of the random forest model.

To deeply analyze the performance of the logistic regression model, we also plotted the heat map of the confusion matrix. The confusion matrix shows the prediction of the logistic regression model on two categories (having diabetes and not having diabetes). By looking at the confusion matrix, we find that the logistic regression model performs well in predicting patients with diabetes, and most of the patient samples are correctly categorized. However, the logistic regression model had more mispredictions in predicting non-diabetic patients than the random forest model, i.e., non-diabetic patients were incorrectly predicted as diabetic patients relatively more often.

Although the logistic regression model was slightly less accurate than the random forest model, it still demonstrated good predictive performance. The advantages of the logistic regression model are its simplicity, ease of interpretation, and computational efficiency. In practical applications, a suitable model can be chosen for diabetes prediction according to specific needs and resource constraints.

4. Conclusion

By comparing the experimental results of the random forest model and logistic regression model, we can conclude the following. The random forest model achieves higher accuracy on the diabetes prediction task, which demonstrates its strong classification ability and its ability to capture nonlinear relationships. The logistic regression model, although slightly less accurate, is simple, easy to understand, and computationally efficient, and may be more suitable for rapid deployment and application in some scenarios. The analysis of the confusion matrix reveals the difference in the performance of the two models on different categories, which provides a direction for the optimization and improvement of the models.

Disclosure statement

The author declares no conflict of interest.

References

[1] World Health Organization, 2013, Global Action Plan for the Prevention and Control of Noncommunicable Diseases 2013–2020. Geneva: WHO.

[2] Erdogan A, Duzgun AP, Erdogan K, et al., 2018, Efficacy of Hyperbaric Oxygen Therapy in Diabetic Foot Ulcers Based on Wagner Classification. The Journal of Foot and Ankle Surgery, 2018, 57(6): 1115–1119.

[3] Wang L, Peng W, Zhao Z, et al., 2021, Prevalence and Treatment of Diabetes in China, 2013–2018. JAMA, 326(24): 2498–2506

[4] Tuomilehto J, Lindstrm J, Eriksson JG, et al., 2001, Prevention of Type 2 Diabetes Mellitus by Changes in Lifestyle Among Subjects with Impaired Glucose Tolerance. New England Journal of Medicine, 344(18): 1343–1350. https://doi.org/10.1056/NEJM200105033441801

[5] Hayes C, Kriska A, 2008, Role of Physical Activity in Diabetes Management and Prevention. Journal of the American Dietetic Association, 108(4): S19–S23.

[6] Salih MS, Khalil R, Zeebaree SRM, 2024, Diabetic Prediction Based on Machine Learning Using PIMA Indian Dataset. Communications on Applied Nonlinear Analysis, 31(5s): 138–156. https://doi.org/10.52783/cana.v31.1008

[7] Naz H, Ahuja S, 2020, Deep Learning Approach for Diabetes Prediction Using PIMA Indian Dataset. J Diabetes Metab Disord, 19(1): 391–403. https://doi.org/10.1007/s40200-020-00520-5

[8] Glasgow RE, 1995, A Practical Model of Diabetes Management and Education[J]. Diabetes Care, 18(1): 117–126.

[9] Garber AJ, Abrahamson MJ, Barzilay JI, et al., 2013, AACE Comprehensive Diabetes Management Algorithm 2013. Endocrine Practice: Official Journal of the American College of Endocrinology and the American Association of Clinical Endocrinologists, 19(2): 327–336.

[10] Watkins PJ, Amiel SA, Howell SL, et al., 2003, Diabetes and Its Management. John Wiley & Sons, United Kingdom.